علم اللُّغة والتِّقنيات المُعاصِرة

د.المُعتزّ بالله السَّعيد

moataz@cu.edu.eg

مرحبا بكم

welcome

29 October 2014

المُحاضَرة السَّادسة

في المُدَوَّنات اللُّغَوِيَّة

٢

29 October 2014

- المدَوَّنات اللُّغَوِيَّة العربيَّة -

- أنواع المدَوَّنات اللُّغَوِيَّة -

29 October 2014

- المدَوَّنات اللُّغَوِيَّة العربيَّة -

- أنواع المدَوَّنات اللُّغَوِيَّة -

29 October 2014

المدَوَّنات اللُّغَوِيَّة العربيَّة

مُدَوَّنة "نايمَيخِن" (NIJMEGEN Corpus)

- أنجَزَها فريقٌ بحثيٌّ في جامعة نايمَيخِن الهولندِيَّة في عامَيْ ١٩٩٥ و ١٩٩٦، بإشراف المعجميّ الهولندي "يان هوخلاند Jan Hoogland"؛ وهي مُدَوَّنةٌ لُغَوِيَّةٌ مكتوبةٌ، جُمِعَت مادَّتُها من النُّصُوص العربيَّة المنشورة على صفحات الويب، بالإضافة إلى الصُّحُف والمجلاَّت ونُصُوص الآداب العربيَّة، وتضُمُّ ما يزيدُ على مليُونَي كلمة. وتُستَخدَمُ هذه المدوَّنة في صناعة معجمين لُغَوِيَّيْنِ ثنائيَّيْن، أحدُهُما (هولنديّ / عربيّ)، والآخَر (عربيّ / هولنديّ).

المدَوَّنات اللُّغَوِيَّة العربيَّة

المُدَوَّنة العربيَّة (CLARA Corpus)

- أنجَزَها فريقٌ بحثيٌّ بمعهد دراسات الشرق الأدنى التَّابع لجامعة تشارلز التِّشِيكِيَّة في عام ١٩٩٧م؛ وهي مُدَوَّنةٌ لُغَوِيَّةٌ مكتوبةٌ، جُمِعَت مادَّتها من الدَّوريَّات العلمية والصُّحُف العربيَّة، وتضُمُّ خمسين مليون كلمة. تُستَخدَمُ هذه المادَّةُ لأغراض الصِّناعة المعجَمِيَّة. وتُعَدُّ المدوَّنة جُزءً من مشروعٍ علميٍّ كبيرٍ ترعاهُ وزارة التَّعليم بالتِّشيك لإنجاز مجموعة من المعجمات الثُّنائيَّة.

المدَوَّنات اللُّغَوِيَّة العربيَّة

المُدَوَّنة العربيَّة (LEUVEN Corpus)

- أُنجِزَت بجامعة لوفان الكاثولِيكِيَّة في بلجيكا بين عامَيْ ١٩٩٥ و ٢٠٠٤؛ وتتنوَّع مادَّتها بين المكتوب والمنطوق: فالمادَّة المكتوبة مستقاةٌ من الصُّحُف والمجلات وكُتُب تعلُّم العربية وتضُمُّ ثلاثة ملايين كلمة، والمادَّة المنطوقة مُستَمَدَّةٌ من الإذاعات العربيَّة والمسرَحِيَّات وتشتمل على ٧٠٠ ألف كلمة. صُنِعَت هذه المدَوَّنةُ ليُستَفادَ منها في بناء مُعجمٍ عربيّ/هولنديّ، يُلَبِّي حاجةَ مُتَعلِّمي العربيَّة من أبناء هولندا وبلجيكا.

المدَوَّنات اللُّغَوِيَّة العربيَّة

المُدَوَّنة إيجيبت (EGYPT Corpus)

- وضعَها مركزُ مُعالَجة اللُّغة في جامعة جون هوكنز John Hopkins في عام ١٩٩٩. وهي مُدَوَّنةٌ لُغَوِيَّةٌ مكتوبةٌ ومُتوازية، مادَّتُها القُرآن الكريم وترجمة معانيه إلى الإنجليزِيَّة والفرنسِيَة، وتُصاحبُها بعضُ الإحصاءات الَّتي أُجرِيَت على نُصُوص القُرآن الكريم. تُستَخدَمُ لأغراض التَّرجمة الآليَّة.

المدَوَّنات اللُّغَوِيَّة العربيَّة

المُدَوَّنة التَّاريخيّة للُّغة العربيَّة (HAC Corpus)

- صَنَعَها الباحثُ (المعتزّ بالله السَّعيد) ضِمنَ الأطروحة الَّتي تقدَّمَ بها إلى جامعة القاهرة للحُصُول على درجة الدُّكتوراه في عام ٢٠١١م. جُمِعَت نُصُوصُ المدَوَّنة – في إصدارها الثَّاني - من التُّراث العربيّ المكتوب عبرَ العُصُور الأدبِيَّة للعربيَّة بدءً من عام ١٥٧ وحتَّى ٢٠١٤م، ويربو عدد كلماتها على مليار ومِئتَي مليون كلمة. أتيحَت هذه المدوَّنة لأغراض البحث العلميّ، برعاية المنظَّمة العربيَّة للتَّربية والثَّقافة والعُلُوم ALECSO.

المدَوَّنات اللُّغَوِيَّة العربيَّة

مجموعة مُدَوَّنات (LDC)

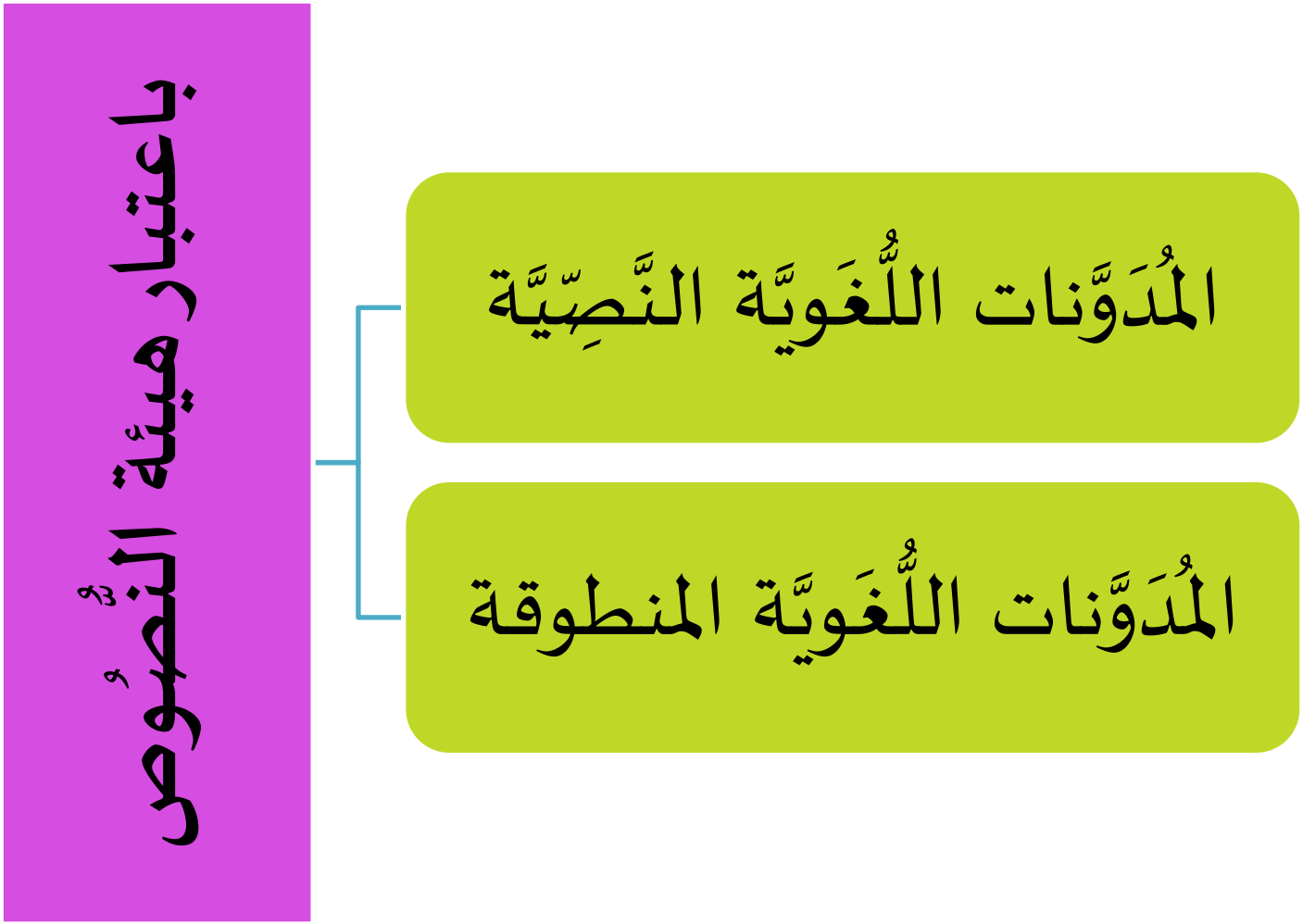
المدَوَّنة	البيانات	التَّطبيقات
Arabic Gigaword	مكتوبة	استرجاع المعلومات
Arabic Treebank	مكتوبة	استرجاع المعلومات، ومُعالَجة اللُّغات الطَّبيعيَّة
الإخباريَّة المتَوازِية	مكتوبة	استرجاع المعلومات، وتعليم اللُّغات، والتَّرجمة الآليَّة
الإخباريَّة المنطوقة	منطوقة	التَّعلُّم الآليّ، والتَّرجمة الآليَّة

- المدَوَّنات اللُّغَوِيَّة العربيَّة -

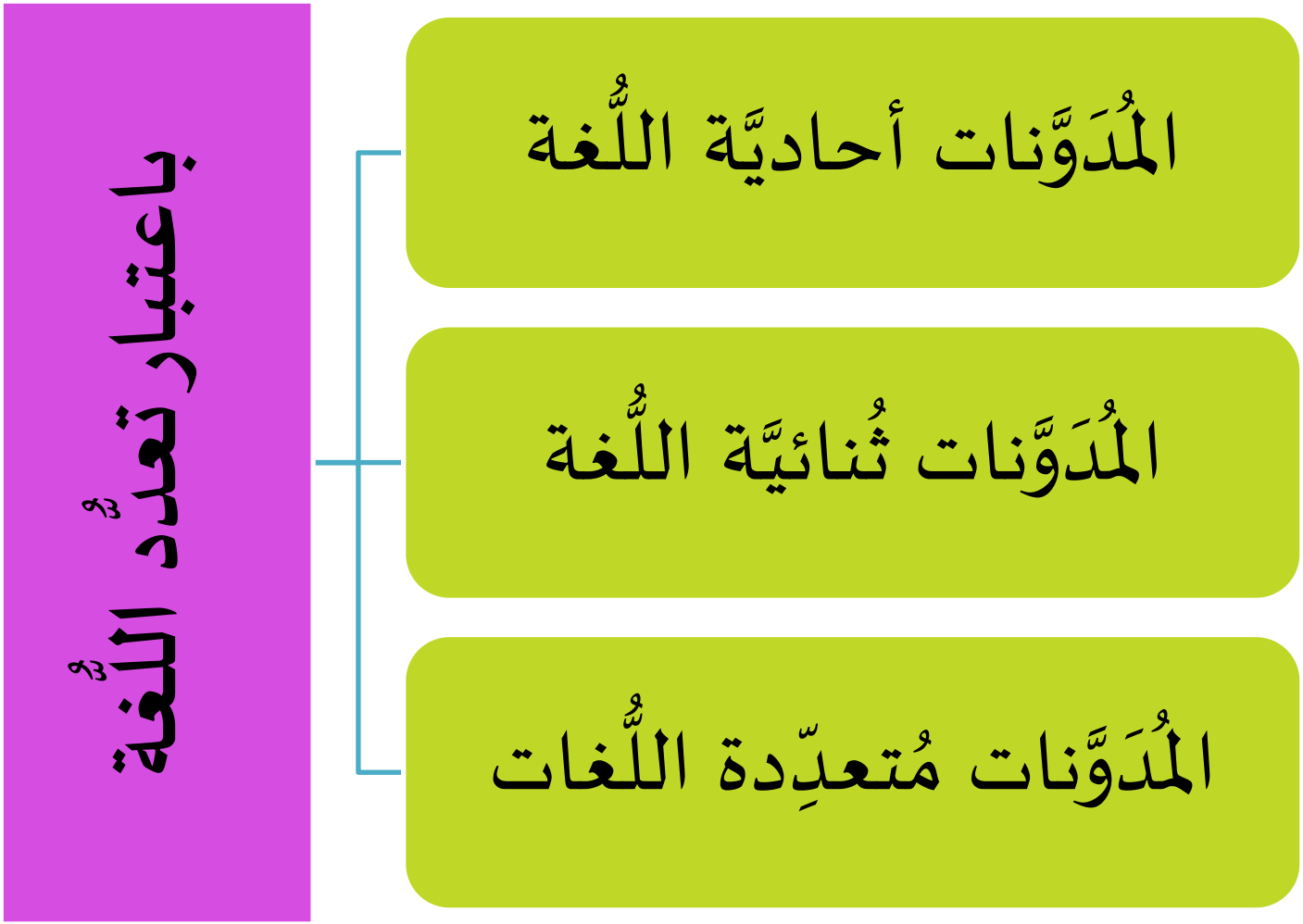
- أنواع المدَوَّنات اللُّغَوِيَّة -

29 October 2014

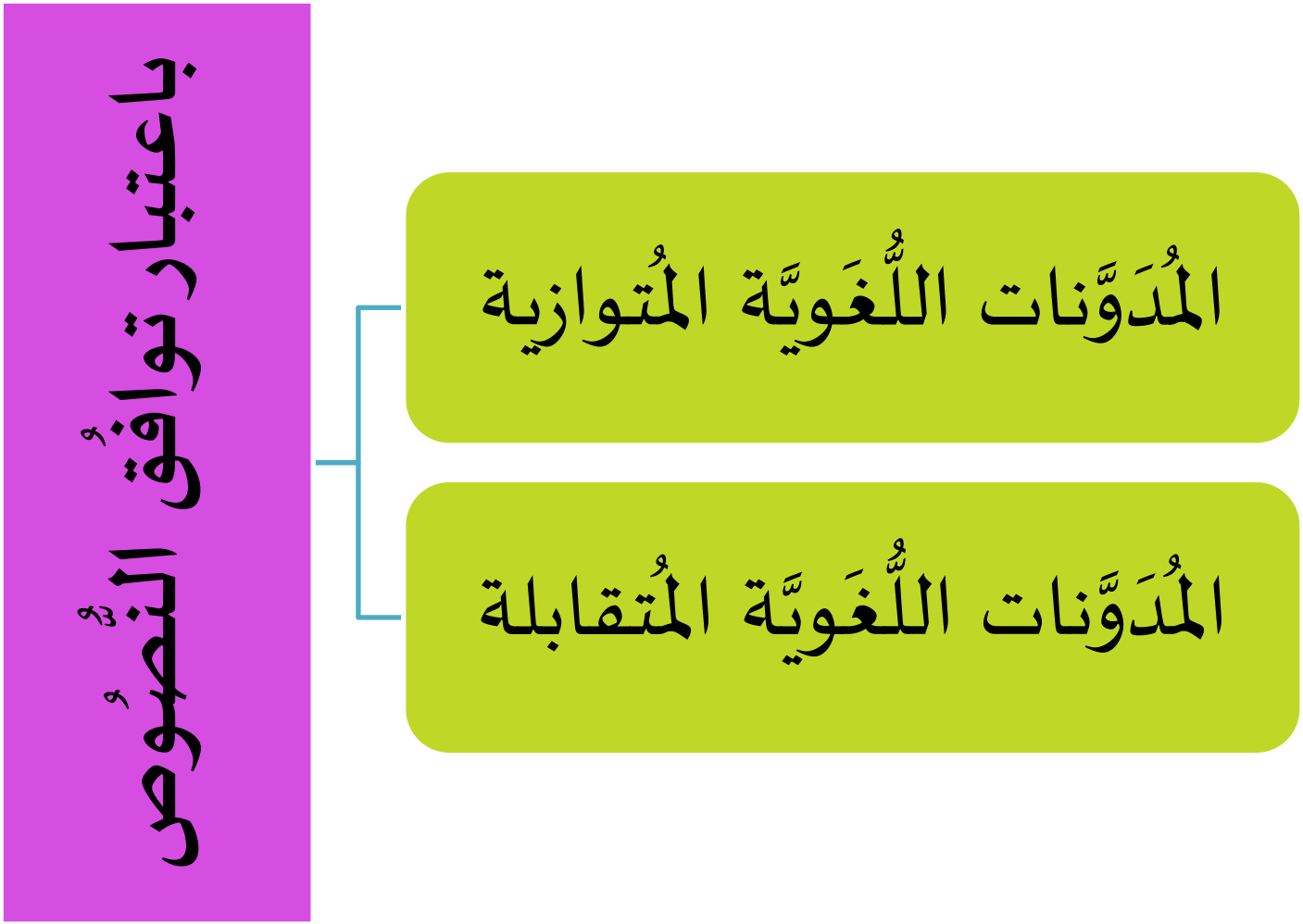
أنواع المدَوَّنات اللُّغَوِيَّة



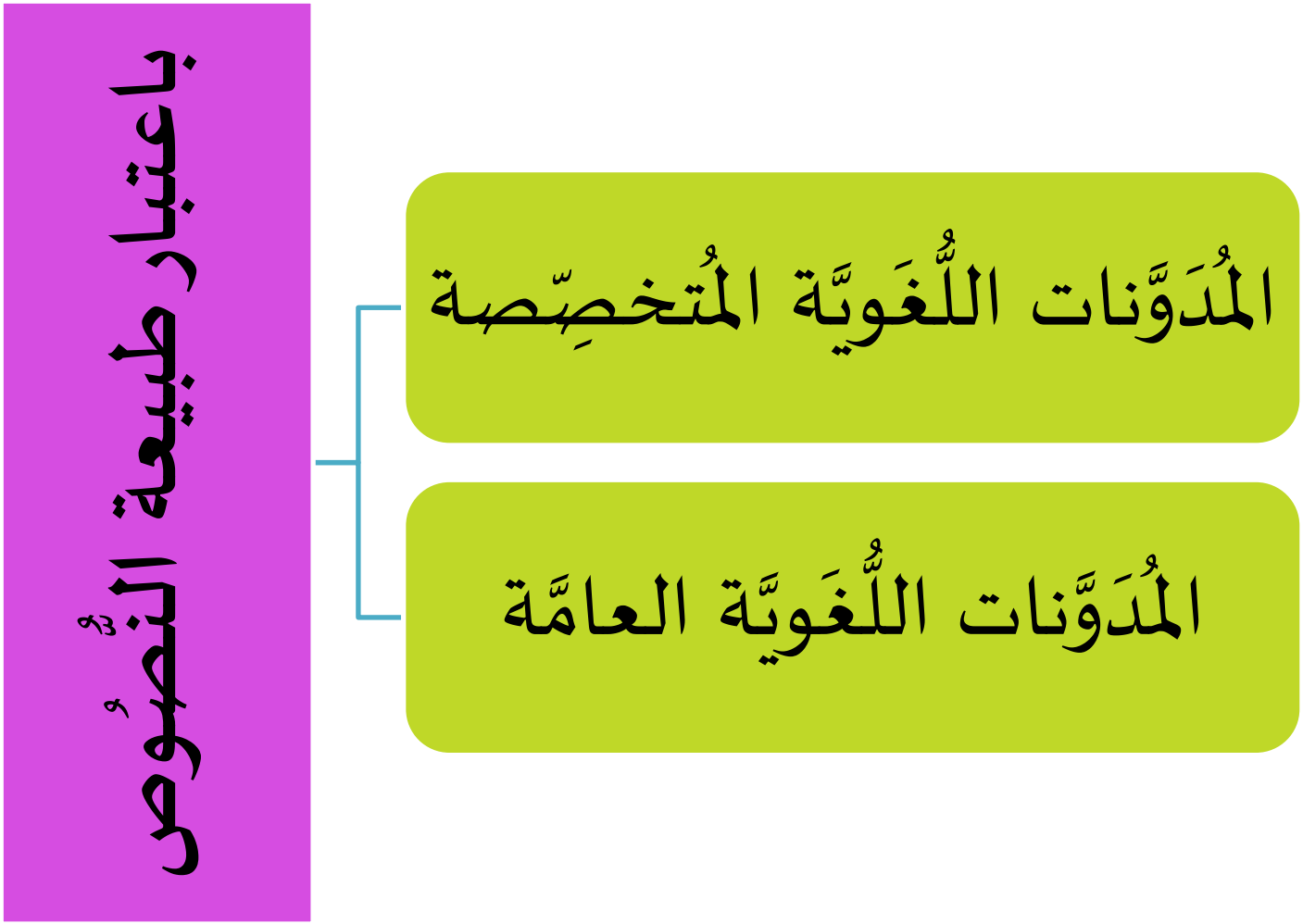
أنواع المدَوَّنات اللُّغَوِيَّة



أنواع المدَوَّنات اللُّغَوِيَّة



أنواع المدَوَّنات اللُّغَوِيَّة



29 October 2014

للتَّواصُل ..

- د. المُعتزّ بالله السَّعيد

moataz@cu.edu.eg

almoataz2006@skype.com

almo3tazbellah@facebook.com

www.ask.fm/AlmoatazBAlSaid

http://scholar.cu.edu.eg/moataz

http://cairo.academia.edu/AlmoatazBAlSaid